Multiple Choice Question Answering in the Legal Domain Using Reinforced Co-occurrence

Jorge Martinez-Gil[1], Bernhard Freudenthaler[1], A Min Tjoa[1,2]

[1]Software Competence Center Hagenberg GmbH
Softwarepark 21, 4232 Hagenberg, Austria

[2]Vienna University of Technology
Favoritenstrasse 9-11/188, 1040 Vienna, Austria
e-mail: name.surname@scch.at

Abstract. Nowadays, the volume of legal information available is continuously growing. As a result, browsing and querying this huge legal corpus in search of specific information is currently a tedious task exacerbated by the fact that data presentation does not usually meet the needs of professionals in the sector. To satisfy these ever-increasing needs, we have designed an appropriate solution to provide an adaptive and intelligent solution for the automatic answer of questions of legal content based on the computation of reinforced co-occurrence, i.e. a very demanding type of co-occurrence that requires large volumes of information but guarantees good results. This solution is based on the pattern-based methods that have been already successfully applied in information extraction research. An empirical evaluation over a dataset of legal questions seems to indicate that this solution is promising.

Keywords: Expert Systems, Legal Information Processing, Knowledge Engineering, Information retrieval, Question answering

1 Introduction

An increasing number of professionals from the legal sector agree that the information explosion concerning national and international legislation makes their work more expensive, tedious and even error-prone. The two major reasons for that are: a) national and international legislation is usually formatted in an unstructured way, and b) the huge volume and speed at which legislation is published usually lead to information overload in their daily activities.

In this context, working with information concerning legislation and case law has always been attractive to computer scientists and practitioners looking for applying for the latest advances on language and semantic technologies. In fact, these technologies have proven to be very useful for solving a number of problems that have traditionally affected the field of legal information processing. In practice, the daily work of these professionals requires reading a large amount of legal material necessary to identify the relevant documents and to identify the correct

fragment that they need. One step in the evolution towards the improvement of these processes come from a subfield from the information retrieval (IR) field, and it is called Question Answering (QA) systems. In fact, the design of systems of this kind is presented as an alternative to overcome the traditional processes by trying to provide accurate and understandable answers to specific questions, rather than presenting the user with a list of search-related documents [11].

In the particular case of the legal domain, the research community agrees that a system allowing to generate automatic responses to legal questions could have a strong impact with a lot of practical implications in their daily activities. The degree of usefulness is such that even the reduced version of the problem that we are addressing here (multiple choice, i.e. responding in a scenario where the answers are already given beforehand [1]) can also significantly help to reduce the workload. This is mainly because a QA system would be able to automatically process a huge amount of legal resources to answer a question or doubt in a matter of seconds, and that means that it could save resources in the form of effort, money and time to many professionals in the legal sector.

To tackle this problem, we have focused on computational techniques for co-occurrence analysis. Techniques of this kind have been widely used in various forms of research on content analysis, text mining, thesauri building, and ontology learning. Here, we propose a specific kind of co-occurrence, i.e. reinforced co-occurrence that it is intended to order to discover latent patterns on huge text corpora. And although our field of application in this work is the legal field, some of the conclusions that are drawn can be extrapolated to a wide range of specific domains. Therefore, with this idea in mind, we present here our research from which the following contributions can be highlighted:

- We propose a new method for the automatic answer of multiple choice questions of legal content based on the idea of computing reinforced co-occurrence.
- We have compiled a dataset of legal questions so that the researchers can try and compare their own solutions, and we have empirically evaluated our approach using the aforementioned dataset of legal questions.

The remainder of this work is organized in the following way: Section 2 reports the state-of-the-art on question answering methods and tools that have proven to be successful in the legal domain. Section 3 presents the fundamentals of our contribution concerning the computation of the reinforced co-occurrence over huge corpora. Section 4 reports the empirical evaluation of our novel approach over a legal dataset and the analysis of the results that we have achieved. Finally, we outline the conclusions and future lines of research.

2 State-of-the-art

A QA system is a kind of computer system intended to automatically reply questions by analyzing different sources of either structured or unstructured information. These sources are usually called Knowledge Bases (KBs). In this context,

there are basically two different approaches to tackle the problem depending on the KBs to be exploited: working with structured KBs, or working with unstructured KBs. Each of them has different advantages and disadvantages. For example, working with structured KB allows exploiting the knowledge represented by using the so-called inference engines, in order to infer new knowledge and to answer questions. However, at present, there is not an automatic way to introduce a new entity into the KB nor to determine with which existing entities should be related and how [15]. Therefore, finding practical solutions is considered as an important research challenge and its matter of intense research [10].

The fact is that not easy to implement these systems, so they have been progressively replaced by another type of more efficient systems based on lighter knowledge models such as knowledge graphs [6] and other enhanced lexical semantic models [22], but in general, it is widely assumed that building a KB is expensive in terms of resource consumption, it is subject to many errors, it is usually difficult and expensive to maintain, and last but not least, a structured KB is usually hardly reusable.

In contrast, IR systems have more practical benefits as most of them have been specifically designed to efficiently process huge amounts of textual data (usually represented in natural language). These huge amounts of data come from existing documents, databases, websites, and so on. For this reason, the most frequent type of QA system that is mentioned in the literature is the one that uses unstructured KBs including different collections of unstructured natural language KBs. In fact, the current generation of QA systems has evolved to extract answers from a wide range of different plain machine-readable resources. These QA systems exploit the massive set of unstructured information available on some sources to retrieve information about any particular question. It is important to note that these QA systems are possible mainly due to recent advances in the big data and natural language technologies. Moreover, since these novel QA systems are capable of processing questions about different domains and topics, they are now used in a wide range of different scenarios [14].

In this context, IR-based solutions represent words in the form of discrete and atomic units. For example, the first approach (and the simplest) could be to query the number of Google results for a specific question and a given answer together. However, this solution has brought a number of problems like the lack of context. To overcome these problems, word processing models such as LSA [5] and term frequency-inverse document frequency (tf-idf) partially solve these ambiguities by using terms that appear in a similar context based on their vector representation, and then they group the semantic space into the same semantic cluster. In this context, one of the best-known QA systems is IBM Watson [8], that it is very popular for its victory in the televised show *Jeopardy* [9]. Although in recent times, IBM Watson has become a generic umbrella that includes other business analytics capabilities.

If we focus strictly on the legal field, we find that QA technology has been very little used in real information systems, and especially in knowledge man-

agement systems [2]. The logic behind these systems is that given a legal issue, the extraction of relevant legal resources and the decision whether or not to use that content to answer the question are two key steps in building a system. In recent times, a number of works have been presented in this context. There are two major branches, a) with structured KB. For example, Lame et al. [12] and Fawei et al. [7] using ontologies, or Xu et al. [21] by exploiting other KBs such as Freebase. And b) exploiting unstructured KBs. For example, Brueninghaus and Ashley with a classical IR approach [4], Bennet et al. with strong focus on scalability [2], Maxwell and Schafer paying attention to context [16], Mimouni et al. with the possibility to make use of complex queries [17], or most modern deep learning techniques from Marimoto et al. [18] and Nicula et al. [19], the latter with good results, although with issues concerning the interpretability of the results.

3 Multiple Choice Question Answering Using Reinforced Co-occurrence

To overcome the current limitations of exiting QA approaches in the legal domain, we propose to automatically analyze co-occurrence patterns belonging to different corpora of unstructured text. Therefore, our approach is intended to automatically process huge amounts of legal information in order to look for evidence allowing to infer the most promising answers with regards to the huge range of questions that the legal professionals could potentially make. In this way, our contribution is a novel approach for automatically answering multiple choice questions concerning a wide range of topics belonging to the legal domain. This approach needs to fulfill two stages: first, we need to calculate alignment matrices between the question and the possible choices using textual corpora, and then we need to normalize the results in order to produce a final outcome and associated ranking of possible answers.

It is not difficult to see that the design of such as text mining approach in this context is far from being trivial. However, our experience in rapid prototyping and testing text mining solutions has shown us that it is possible to reach a reasonable level of success [14]. According to our experience, the solution that works best is a method with four levels of co-occurrence depending on the context whereby the question and the choice being evaluated can be found together.

On the other hand, the problem that we are addressing here is based on short answer models. The reason is that these models provide the potentially correct answer in the form of a number, a name, a date, or even a short phrase or text fragment. This makes the work of our text mining engine easier. It is also important to note, that this assumes that there are different ways of asking questions, and most of them are characterized by the formulation of questions expressed by interrogative particles (i.e. what, who, why, when, where, where) or some kind of is-a association. At the same time, the aforementioned possible choices are expressed in natural language.

Although the concept seems to be not to difficult to understand, there are huge technical limitations for its development from a pure engineering perspective. In fact, this approach is limited by an important number of technical issues which should be overcome. These limitations, originally identified by [3], are inherent to the process of massively text mining. In order to facilitate overcoming these limitations, our system is designed in the form of a pipeline, i.e. a workflow whereby the data flow into processes so that the output of one process is the input of the next one. Figure 1 shows us an overall view of our IR pipeline. These components are related to each other and process the textual information available on different levels until the QA process has been completed. The natural language questions formulated to the system are processed initially by the question analysis component.

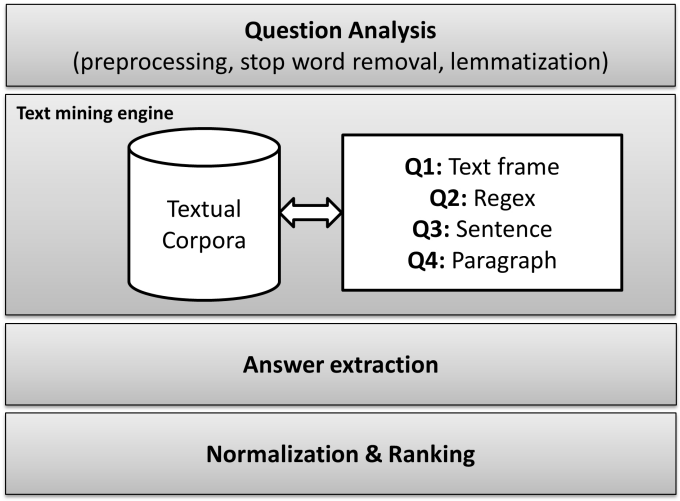

Fig. 1: Pipeline designed to answer the multiple choice tests. First of all, questions and answers need to be pre-processed. After this, a text mining engine is in charge of mining reinforced co-occurrence patterns. Then, these patterns are analyzed. Finally, the results are normalized and a ranking of potential choices is provided

Then, the system continues working by conveniently dividing the information into different parts which will be transferred to the following process which is a text mining engine that looks for the reinforced co-occurrence of the question and each possible answer. Then, the answer extraction module compiles the raw data resulting from the mining phase. Finally, it is necessary to normalize the raw data and create the final ranking to be delivered. The main modules of our QA system could be summarized in the following steps:

– *Question Analysis.* It is in charge of pre-processing both the question and the possible answers. To do that, it is necessary to remove the stop words,

to proceed with a lemmatization process, i.e. determining what is the root of the words so irregular forms (e.g. plurals, third person, etc...) does not affect to the co-occurrence, and remove very common adjectives and verbs.
- *Reinforced Co-Occurrence Calculation.* It consists of calculating how many times the pre-processed question and the evaluated answer co-occur together in the same text frame, in the same text expression, in the same sentence, and in the same paragraph.
- *Answer Extraction.* It consists of compiling the results and assign them to each of the possible choices. After this process, we have just raw values that need to be refined.
- *Answer Normalization and Ranking.* Normalization is the process of mitigating the impact of the outliers on the final decision. In this work, we usually work with exponential reductions, but other methods need to be considered in future work. Ranking consists of creating an ordered list of response according to the score obtained after normalization.

3.1 Running Example

In order to illustrate how our approach works, we have designed a running example to better understand how our pipeline processes the information. Let us think in a question whereby we would like to know the kind of document represented by the European Convention on Human Rights. Let us think how the question could be, and how the different choices would look like.

```
What document is the European Convention on Human Rights?
a) A statute
b) Delegated legislation
c) An EU directive
d) A treaty (correct choice)
```

Then, our system would start by evaluating the suitability of the first answer, i.e. statute. To do that, we can see in Figure 2 the graphical summary of how this process is performed: The question and the associated choices have to be preprocessed in order to remove non-relevant words, perform lemmatization, etc. Then, this information has to be submitted to text mining engine, where a dispatcher tries to look for the reinforced co-occurrence of the pair question-possible answer by scanning all possible co-occurrences within the corpora. As a result, we get the reinforced co-occurrence values that have to be normalized so the outliers might not have an extreme weight in the final value.

After repeating this process for each of the possible choices, we have that in this case, our solution must discern whether it is about a statute, a delegated legislation, an EU directive or a treaty. In Table 1, a normalization has been applied. In this case, normalization consists of gradually reducing the value associated with the co-occurrence of very general terms since this adds an excessive noise at the time of obtaining a meaningful response. In this case, statute and document have a high degree of co-occurrence that can make other parts of the

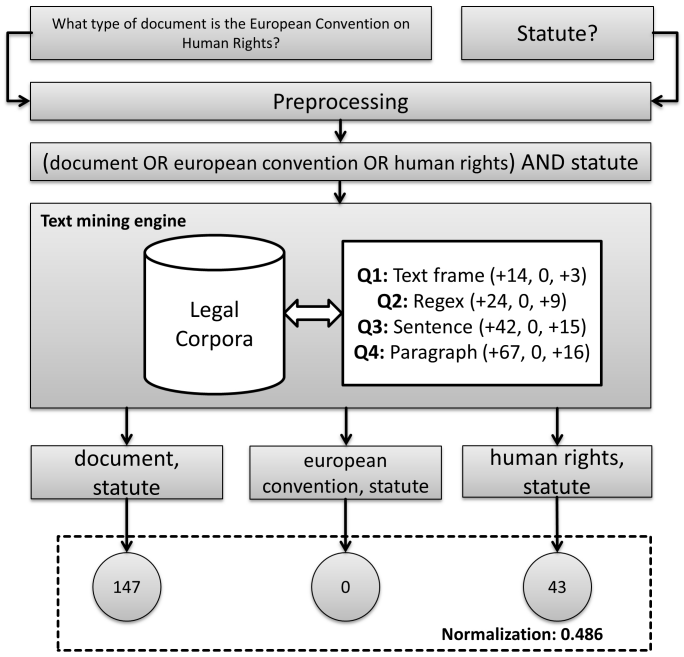

Fig. 2: View of one iteration whereby a question and a potential answer are evaluated

question lose prominence, so we must proceed to reduce the impact that such co-occurrence has on the final result.

Therefore, the choice that our system would select as the correct one is d) A treaty, what is also the correct one according to the ground truth. The second would be a) A statute. And the other two possible choices has no reinforced co-occurrence, so they would not even be considered as possible answers. Additionally, a heatmap allows to visually inspect the rationale behind the result. This is mainly due to the fact that in some scenarios requiring accountability and/or interpretability, it is not just enough to provide the answer, but also some reasons for helping to interpret that answer.

4 Results

We explain here the results. It is important to remark that these results are highly dependent on the base corpus that will be processed. Choosing a relevant, specific base corpus to evaluate each of the possible choices is really important in this context. On the other hand, the task of evaluating the system is of vital importance, as it will assess the performance, as well as the accuracy of the techniques. In this work, we have chosen the strictest methodology to evaluate systems, which consists of binary classification: the answer was right or wrong.

	statute	delegated legislation	EU directive	treaty
document	0.376	0.000	0.000	1.000
european convention	0.000	0.000	0.000	0.000
human rights	0.110	0.000	0.000	0.441
Final score	0.486	0.000	0.000	1.441
Ranking	2	-	-	1

Table 1: Normalized results obtained for the reinforced co-occurrence, final score and ranking proposal

4.1 Solving the benchmark dataset

We show here the results that we have obtained when testing our solution. The dataset has been generated by picking randomly legal questions suggested from a number of books from the Oxford University Press[1]. Moreover, it is important to remark that the techniques are applied automatically over the dataset of questions without prior knowledge of the answers (e.g. without machine learning). After comparing the answers with the results, the performance is determined through the accuracy metric.

Therefore, we have obtained 13 correct answers from a total of 20 questions. This means that we got a 65% of accuracy. Table 2 shows a comparison with other approaches. By responding randomly there is a 25% chance of guessing the correct answer, so that is the score we have established as baseline. Please note that, for the sake of fair comparison[2], we just include approaches without machine learning capabilities (as ours). At the same time, we hope that the compilation of this dataset will stimulate the evaluation of more QA systems.

Approach	**Correct Answers**	**Accuracy**
Baseline	5	25%
Calcipher [20]	7	35%
Li et al. [13]	9	45%
LSA-Classic [5]	9	45%
Our Approach	13	65%

Table 2: Comparison with other approaches

QA technology is becoming a very important solution in a wide range of areas overloaded by the constant generation of large amounts of information. In this context, being able to automatically answering specific questions in a correct manner can contribute to alleviating the problem of dealing with those huge amounts of data. Our approach is able to offer good results, at an affordable

[1] http://global.oup.com

[2] Although we foresee learning the parameters of our system as future work

cost (in terms of money, time, and effort needed), without the need for training, and with great facilities for interpretability. This technology, however, faces some obstacles in its development related to the amount of engineering work to properly tune the parameters involved along the IR pipeline.

5 Conclusions and Future Work

In the context of the legal domain, methods and techniques for answering specific questions are in high demand, and as a result, a number of solutions for QA have been developed to respond to this need. The major reason for that is that the capability to automatically answer questions by means of computers could help alleviate a problem involving tedious tasks such as an extensive information search what is, in general, time-consuming. By automatically providing hints concerning a wide number of legal topics, lots of resources in the form of effort, money and time can be preserved. In this work, we have presented our research on automatically addressing multiple choice questions and the development of techniques for automatically finding the correct answer by means of IR pipeline that implements reinforced co-occurrence.

We have seen that although approaches based on structured KB often yield good results, it is difficult to use them in practice mainly due to the cost when building such structured KB (i.e. it is expensive in terms of effort, money and time needed) and it is often very difficult to find experts with enough knowledge for curating the KB. In contrast, our approach has a number of practical benefits when selecting the actual right answer from a list of the possible answers due to the advances in big data processing and natural language technology. Moreover, in the present work, we have not yet fully explored the characteristics of legal texts in order to utilize these features for building our legal QA system. In fact, properties such as references between documents or structured relationships in legal statements should be investigated more in depth as part of future work.

As additional future lines of research, we also need to work towards overcoming a number of technical limitations. This includes the capability to work with different multilingual corpora at the same time, the proper processing of verbs when formulating questions and evaluating potential answers, and the proper tuning of the different system parameters by means of a training phase. We think that by successfully addressing these challenges, it is possible to build solutions that can help the legal practitioners to overcome one of the most problematic issues that they have to face in their daily work.

Acknowledgements

This research work has been supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the State of Upper Austria in the frame of the COMET center SCCH.

References

1. B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, M. Demirbas: Crowdsourcing for Multiple-Choice Question Answering. AAAI 2014: 2946-2953.
2. Z. Bennett, T. Russell-Rose, K. Farmer: A scalable approach to legal question answering. ICAIL 2017: 269-270.
3. S. Blohm, P. Cimiano: Using the Web to Reduce Data Sparseness in Pattern-Based Information Extraction. PKDD 2007: 18-29.
4. S. Brueninghaus, K. D. Ashley: Improving the representation of legal case texts with information extraction methods. ICAIL 2001: 42-51.
5. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, R. A. Harshman: Indexing by Latent Semantic Analysis. JASIS 41(6): 391-407 (1990).
6. J. Ding, Y. Wang, W. Hu, L. Shi, Y. Qu: Answering Multiple-Choice Questions in Geographical Gaokao with a Concept Graph. ESWC 2018: 161-176.
7. B. Fawei, J. Z. Pan, M. J. Kollingbaum, A. Z. Wyner: A Methodology for a Criminal Law and Procedure Ontology for Legal Question Answering. JIST 2018: 198-214.
8. D. A. Ferrucci: Introduction to This is Watson. IBM Journal of Research and Development 56(3): 1 (2012).
9. D. A. Ferrucci, A. Levas, S. Bagchi, D. Gondek, E. T. Mueller: Watson: Beyond Jeopardy! Artif. Intell. 199-200: 93-105 (2013).
10. K. Hoeffner, S. Walter, E. Marx, R. Usbeck, J. Lehmann, A.-C. Ngonga Ngomo: Survey on challenges of Question Answering in the Semantic Web. Semantic Web 8(6): 895-920 (2017).
11. O. Kolomiyets, M.-F. Moens: A survey on question answering technology from an information retrieval perspective. Inf. Sci. 181(24): 5412-5434 (2011).
12. G. Lame: Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations. Artif. Intell. Law 12(4): 379-396 (2004).
13. Y. Li, D. McLean, Z. Bandar, J. O'Shea, K. A. Crockett: Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Trans. Knowl. Data Eng. 18(8): 1138-1150 (2006).
14. J. Martinez-Gil, B. Freudenthaler, T. Natschlaeger: Automatic recommendation of prognosis measures for mechanical components based on massive text mining. IJWIS 14(4): 480-494 (2018).
15. J. Martinez-Gil: Automated knowledge base management: A survey. Computer Science Review 18: 1-9 (2015).
16. K. T. Maxwell, B. Schafer: Concept and Context in Legal Information Retrieval. JURIX 2008: 63-72.
17. N. Mimouni, A. Nazarenko, S. Salotti: Answering Complex Queries on Legal Networks: A Direct and a Structured IR Approaches. AICOL 2017: 451-464.
18. A. Morimoto, D. Kubo, M. Sato, H. Shindo, Y. Matsumoto: Legal Question Answering System using Neural Attention. COLIEE@ICAIL 2017: 79-89.
19. B. Nicula, S. Ruseti, T. Rebedea: Improving Deep Learning for Multiple Choice Question Answering with Candidate Contexts. ECIR 2018: 678-683.
20. M. Stam: Calcipher System. Retrieved from https://github.com/matt-stam/calcipher, on 01-04-2019.
21. K. Xu, S. Reddy, Y. Feng, S. Huang, D. Zhao: Question Answering on Freebase via Relation Extraction and Textual Evidence. ACL (1) 2016.
22. W.-T. Yih, M.-W. Chang, C. Meek, A. Pastusiak: Question Answering Using Enhanced Lexical Semantic Models. ACL (1) 2013: 1744-1753.